REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	2/14/97	FINAL TECHNICAL REPORT 10/1/96-12/31/96

4. TITLE AND SUBTITLE
THE MATHEMATICS OF INFORMATION CODING, EXTRACTION AND DISTRIBUTION

5. FUNDING NUMBERS

6. AUTHOR(S)
AVNER FRIEDMAN

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Mathematics and Its Applications
University of Minnesota
514 Vincent Hall
206 Church Street SE
Minneapolis, MN 55455

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In the last decade, considerable engineering progress has been made in the technology of large-scale information systems. We are now at the point where ambitious deployments of global information networks are soon to be realized but serious problems remain in the areas of scalability, modeling and applications. The workshop on "The Mathematics of Information Coding, Extraction and Distribution" brought together experts in various areas of mathematical and practical information theory and modeling to formulate the problems, explore new analytic methods and exchange ideas. We addressed applications areas such as "data mining", compression, database theory and machine learning, with special attention to the interactions between these areas from the analytical and mathematical points of view. The workshop devoted half time to dissemination of new technical results and half time to the formulation of new paradigms and problems for future research. Proceeding of the workshop is under preparation and will appear in the IMA Volumes in Mathematics and its Applications Series to be published by Springer–Verlag, New York . Currently, submissions are being edited/refereed by George Cybenko, Dianne O'Leary, and Jorma Rissanen.

14. SUBJECT TERMS
information coding, extraction, distribution, data mining, compression, database theory and machine learning

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

THE MATHEMATICS Of INFORMATION CODING, EXTRACTION AND DISTRIBUTION

FINAL REPORT

Period Covered by This Report: 10/1/96–12/31/96

AVNER FRIEDMAN

February 14, 1997

OFFICE OF NAVAL RESEARCH

ONR Grant N/N00014-97-1-0074

INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS
514 Vincent Hall
University of Minnesota
Minneapolis, Minnesota 55455

FINAL REPORT

Title of Project: **The Mathematics of Information Coding, Extraction and Distribution**

ONR Grant Number: N/N00014-97-1-0074

Period Covered: 10/1/96–12/31/96

Principal Investigators: Avner Friedman and Robert Gulliver

I. Summary of problem studied.

On November 11–15, 1996 George Cybenko (Dartmouth), Dianne O'Leary (Univ. of Maryland) and Jorma Rissanen (IBM Almaden organized a workshop under the auspices of the Institute for Mathematics and Its Applications (IMA), in Minneapolis, Minnesota, for a workshop called "The Mathematics of Information Coding, Extraction and Distribution."

In the last decade, considerable engineering progress has been made in the technology of large-scale information systems. We are now at the point where ambitious deployments of global information networks are soon to be realized but serious problems remain in the areas of scalability, modeling and applications. The workshop on "The Mathematics of Information Coding, Extraction and Distribution" brought together experts in various areas of mathematical and practical information theory and modeling to formulate the problems, explore new analytic methods and exchange ideas. We addressed applications areas such as "data mining", compression, database theory and machine learning, with special attention to the interactions between these areas from the analytical and mathematical points of view. The workshop devoted half time to dissemination of new technical results and half time to the formulation of new paradigms and problems for future research.

2. Abstracts from the workshop organizers and speakers.

Michael T. Orchard (Princeton University)
Space-frequency methods in image coding

In the past several years, two disjoint clusters of researchers have been exploring the application of time-frequency ("space-frequency" when applied to imagery) analysis tools to image coding. On one hand, image coding researchers have been developing increasingly powerful "wavelet-based" image coding algorithms that currently far outperform non-wavelet methods. On the other hand, a growing community of mathematicians and statisticians have been studying the approximation of spaces of functions using "nonlinear approximation" methods with wavelet basis. Results in this direction have established some fundamental advantages of the wavelet basis compared with the Fourier basis for approximating certain classes of functions.

This talk will try to bridge the gap between these two communities. We begin with an overview of the state-of-the-art in image coding, focusing on the role of wavelets in addressing issues critical to efficient compression. We then sketch some recent mathematical results in nonlinear approximation, and show how the analysis reaches similar conclusions about the important issues in image compression. While the mathematical results characterize asymptotic coding performance, we describe how realistic coding regimes can be identified for which the mathematical results provide good qualitative characterization of performance. We briefly review several state-of-the-art wavelet coding algorithms, and show how their superior performance is well predicted by the mathematical analysis. Finally, looking beyond today's top-performing wavelet algorithms, we offer a simple exercise that points to the potential for future advances.

Robert M. Gray (Stanford University)
Lossy compression, classification, and regression

The theory and design of lossy compression systems share many ideas and techniques with statistical classification and regression and hence also with image segmentation. These similarities motivate incorporating a Bayes risk term into a Shannon source coding formulation in order to model a system combining quantization with either classification (detection) or regression (estimation). This provides some new (and old) algorithms for compression, classification, and regression individually, but more interestingly it provides an approach to the joint optimization of systems involving both compression and classification. Examples include the compression of digital mammograms with built in highlighting of microcalcifications and the compression of image data while simultaneously segmenting the image into different local types for separate rendering or printing. The design of such codes involves ideas from clustering and tree-structured statistical methods and it leads to issues involving the combination of quantization, probability density or mass estimation, and classification and regression. The resulting codes have as "extreme points" universal source codes and classified vector quantizers. This talk will survey the basic ideas, illustrate them with examples, and describe some of the algorithms under current study along with several conjectures about their asymptotic performance.

Ahmed H. Tewfik (Univ. of Minnesota)
Protection, manipulation and storage of digital audio, image and video

Audio, image and video data is increasingly available in digital format. This digital representation facilitates the reproduction, manipulation and transfer of the data. In particular, users can easily search and edit the data. However, it also facilitates theft and unauthorized use of the data. In this talk, I will discuss novel techniques for protecting and authentifying audio, image and video data. I will also describe novel approaches to data encoding that support retrieval by content and flexible playback over different channels and monitors.

Jorma Rissanen (IBM Almaden Research Center)
Universal modeling and coding

Universal models are an outgrowth of the ideas rooted in universal coding, stochastic complexity, and modeling by the shortest code length or the *MDL* principle, which themselves are modifications of algorithmic complexity, as introduced by Solomonoff. Much like algorithmic complexity, which permits the definition of a universal probability model, albeit a noncomputable one, stochastic complexity can be used to define a computable model, which is universal for a class of probability measures. Although it can fully learn to imitate these models from their samples only asymptotically, we can prove for many classes of models that asymptotically no "estimate" of the data-generating machinery can perform better than the universal model, which then may be regarded as the ultimate "estimator".

In reality, such a foundation for statistics does not require the data, let alone the parameters, to be a sample from any population. The selected probability models are just a convenient way via the Kraft inequality to describe good codes without need to resort to the (false) claim that any of them has generated the data. The central issue then becomes model fitting rather than estimation, in which the elusive model complexity plays a natural and essential part. This brings advantages well beyond the reach of the traditional techniques, especially in complex modeling problems.

This talk will outline algorithms for universal density and regression models as well as a new way to do universal coding.

Julia Abrahams (Office of Naval Research)
Minimum discrimination variable length parsing and coding problems

A common minimum-discrimination perspective unifies several variable-length parsing and coding problems along with joint parsing/coding problems or dual-tree coding. In particular, Stubley and Blake's minimum-discrimination parse tree problem corresponds to Karp's variable-length unequal-costs coding problem for arbitrary source distributions in the sense that, where Stubley and Blake minimize the discrimination D(P,Q), Karp minimizes D(Q,P). In the special case that Q is uniform, Stubley and Blake's problem is Tunstall parsing and Karp's problem is Varn coding. In the special case that P is dyadic, Stubley and Blake's problem is of interest because Karp's problem is Huffman coding.

Paul H. Siegel (Univ. of California, San Diego)
Mathematical aspects of Trellis-coded modulation

Trellis-coded modulation (TCM), a coding technique that combines the functions of error control coding and signal modulation, has revolutionized digital communications during the past 15 years. Its greatest success has been achieved in voice-band data modems, where consumer products now offer reliable data transmission at rates approaching information-theoretic limits. Other applications for coded-modulation techniques are emerging in digital information storage and personal wireless communications.

The evolution of TCM has been characterized by a fast-paced and mutually-enriching interplay between practical engineering advances and theoretical progress. This talk will discuss TCM and some of the connections to the mathematical areas of symbolic dynamics, system theory, and the theory of lattices.

Gil Strang (Massachusetts Institute of Technology)
The search for a good basis

An image corresponds to a very long vector, with one component for each pixel (three components for a color image). By a change of basis the long vector is concentrated into a much smaller number of components, ready for compression. We study the block Toeplitz matrix that produces a new basis from a bank of two filters – lowpass and highpass. The filter coefficients determine the success of the compression. They also determine whether iteration of the lowpass filter (with rescaling) will lead to a useful wavelet basis for function spaces.

Thus the construction of wavelets comes from a problem in matrix analysis. Actual compression uses 4–5 iterations of the basis change.

Cynthia Dwork (IBM Almaden Research Center)
Plausible and implausible copyright protection techniques

Copyright protection for digital content — software, still images, video, online databases, text, *etc.* — is emerging as an area of intense interest and activity. It is also an area of intense hype. This talk will discuss several proposed software techniques for enhancing copyright protection and will examine the implicit and potentially unreasonable infrastructure assumptions upon which some of these techniques are based.

George Cybenko (Dartmouth College)
Transportable information agents: Practice and theory

Agent Tcl is a transportable agent programming system developed at Dartmouth. The system allows complete process migration across architectures and uses a variety of messaging constructs. Agent Tcl is being used by Lockheed Martin, Siemens and various others to handle complex distributed information processing tasks. In addition to describing the main features of Agent Tcl, I will present some results dealing with mathematical models describing the control of information processing systems. Those results use simple constructs from stochastic control and information theory.

Chris Atkeson (Georgia Institute of Technology)
Scaling up numerical machine learning

Numerical machine learning algorithms attempt to find structure in data. Approaches range from using parametric models such as neural networks to using non-parametric models. This talk will explore the differences between batch learning applications in which a fixed training set is used and continuous learning in which new data is continuously added to the training set.

Dianne P. O'Leary (University of Maryland)
A Semi-Discrete matrix decomposition for latent semantic indexing in information retrieval

In Latent Semantic Indexing, a set of documents is represented by a matrix whose entries are measures of frequencies of occurrences of terms in each document. This matrix is then approximated by a sum of rank-one matrices determined using a singular value or related decomposition.

In this work we investigate the use of a semi-discrete approximation to the matrix in which each rank-one matrix can be represented by a scalar multiplying two vectors whose entries equal plus or minus one or zero. Compression rates are thus much greater than for other decompositions, and the processing of each query is much faster. We compare the methods on three databases and discuss the effects of various frequency measures.

This is joint work with Tamara Gibson Kolda.

Geoff Davis (Dartmouth College)
Arithmetic channel coding

We describe a simple generalization of arithmetic coding that allows the addition of non-integral amounts of redundancy to a bitstream. The resulting channel codes are formed from binary expansions of points in a Cantor set. We describe a work-in-progress attempt to use stochastic properties of our deterministic algorithm to replace the random codes used in proofs of Shannon's channel coding theorem. Our attempt is motivated in part by work on normal numbers and uniformly distributed sequences. Our preliminary findings provide some intriguing insights into some properties of the Cantor ternary set.

Eric Metois MIT Media Lab
The state of the state

Research in nonlinear dynamics in the last decade has led to a number of much more broadly applicable techniques for inferring the underlying unknown state of a system from accessible observables, and then building predictive models in the recovered state-space. I will discuss the relationship between state reconstruction and signal separation by time-delay embedding, and estimation by conventional linear filters. I will then look at how nonlinear dissipative entrainment can be applied to state estimation in coding problems, and the connection to recursive estimation. The talk will close with a description of Cluster-Weighted Modeling, a new framework for the associated inference problem for nonlinear stochastic data in high-dimensional spaces.

Duncan Buell (IDA/Center for Computing Sciences)
Large problems in information handling

The Center for Computing Sciences is a federally-funded research laboratory working on problems in large-scale computing and mathematics relevant to the U. S. Department of Defense. Many of these problems invole the handling, processing and extraction of information from very large volumes of data. In this talk I will discuss aspects of some of these problems.

Manfred Opper (Weizmann Institute of Science)
Worst-case redundancy and prediction over sequences

We discuss the game of sequentially assigning probabilities to symbols in a sequence. A cumulative logarithmic (entropic) loss may be interpreted as the codelength for the sequence. We give general upper and lower bounds on the cumulative minimax regret (redundany) for two scenarios: In the random case, the data are generated independently from an unknown distribution from a given family and the player tries to minimize the expected loss for the worst distribution in this family. In the worst case over sequence scenario, NO probabilistic assumption about the generation of the data is made. The player compares her loss with the the loss of an adversary who knows the entire sequence in advance and uses always the best distribution from the family for each given sequence. In both cases it is possible to bound the cumulative regret in terms of a metric entropy for the family of distributions.

This is joint work with David Haussler.

3. Publications.

The purposes of the workshop were achieved. A proceedings is currently in the process of preparation, and will be published by Springer-Verlag, New York. The tentative title is "The mathematics of information coding, extraction and distribution." All submissions will be edited/refereed by George Cybenko, Dianne O'Leary, and Jorma Rissanen.